Survey on Privacy Preservation Technique: Data Masking

Brijesh R Patel[1] , Jignesh B. Maheta[2]
M.E in Computer Engineering
GTU PG School, Ahmedabad[12]

Abstract - **In Today's world, preserving privacy of individual's data is one of the main issues for organizations. Organizations do data sharing for business purposes and this leads to breaches in data privacy and security .Due to this organizations are suffering from business losses every year, in whole world. So, there is a serious demand to protect private data with techniques having effective cost and data utility is needed. Many researches are already done in this area. Many techniques like Encryption, Perturbation, k- anonymity, etc. are there to protect data privacy. In this paper I am trying explain all this techniques, which are developed by many authors. And also explain technique Data Masking, which comes due to Ubiquitous environment. Also I am trying to compare Data Masking techniques with these other techniques.**

1. INTRODUCTION

Today every organization follows common procedure of application development life cycle and makes production data available for testing. This data contain sensitive information and it is visible as it is in testing environment. This data can be use by any unauthorized person for doing malicious activities. Though there are legal agreements in organizations but they serve only post-facto measure after identity or privacy of data have already taken place. But in test environment data need to be same as production data, as to check functionality and reliability of applications. In this situation organizations need cost effective solution. Because Cost of managing data is 5-10 times more than initial acquisition cost [1].So, Data masking Techniques [2] comes into the picture. Though there are many technologies but data masking is flexible, provide more privacy and data utility, cost effective and easy to implement. In the first section this paper I will explain basic of Data Masking Technique. In later section I will compare techniques with Data Masking Technique. After this, section contains existing Data masking Techniques and area where Data masking is currently using.

2. DATA MASKING

When production data is ready then mask this data and then pass it to test environment. We can achieve this using Masking Technique. Data masking mask the data in such way that it looks like original Data but in actual it is not. So, Data masking is de-identifying or obscuring specific data within specific data elements within database table or column [3]. In other word, Data masking is the process that transfers the original data to another fictional data that looks like original data and we can use this masked data to gain additional knowledge of sensitive information.

Also the objective of Data Making is that we can use masked data in such a way that it is original data and at the same time we can reduce risk of privacy breaches. So, this keeps privacy and identity of Data private. An unmasked data has highest Data utility, but no privacy. And in another way encrypted Data has privacy but no Data utility. But Data masking Technique combine these both and provide privacy.

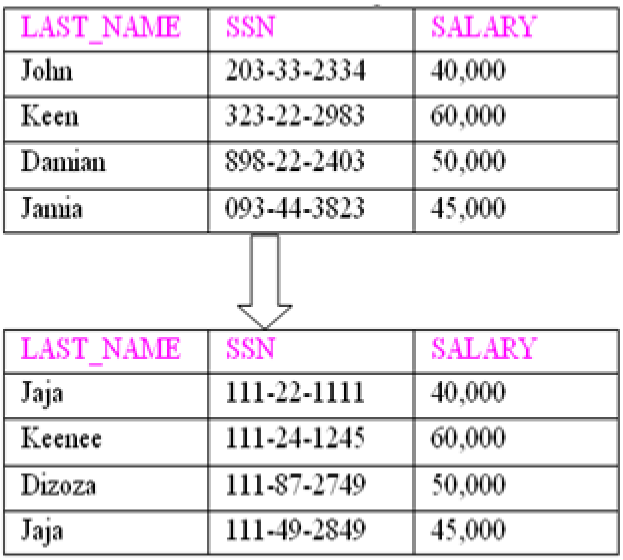
In many areas like Government sector, bank sector, medical sector, location based application, etc. need data available to developer and tester for providing service. But this data contains much sensitive information so it requires privacy protection. If they give original data to developer and tester as it is, malicious activities can be done by them. Data masking technique give the solution to this situation. Data masking masks the data and enables developer and tester to use realistic data and produce valid output.

Data Masking Process: Now, masked data is just data without information. Some government rules and company policies should be followed while doing Data masking. The number of fields to be masked in masking process is very greatly by application and organization requirements. Furthermore there is no single way to the correct masking algorithm or methodology. Indeed more than one technique is used in masking [4][5]. Masking techniques may be unique to every organization as per their requirement. But building of masking technique commonly includes:

1). Simple Masking: This simply replaced data with static set of null values such as XXXX or 9999. This technique is used in simple manual masking process. This technique is not feasible when applications are executed against live data.

2). Numeric Manipulation: This approach simply increment or decrement data by given range. For example data value can be decreased by 15% or balance is increased by 2000. But this simple approach should be used treated by caution, as simple algorithm could deciphered and again the data may no longer represent the production data characteristics.

3). Data Substitution: This is very commonly use approach and it can be extremely effective if it is used well. In this approach data is substituted with an alternative data which is randomly determined or by using replacement mechanism. The integrity of this approach is dependent on data substituted. This approach uses external data sources for data substitution, like birth day is substituted by other valid zip code of town, etc. Whatever techniques are, it is critical to define appropriate action for sensitive data element and to be able to apply a consistent masking process which is propagated through all related data success [6].



LAST_NAME	SSN	SALARY
John	203-33-2334	40,000
Keen	323-22-2983	60,000
Damian	898-22-2403	50,000
Jamia	093-44-3823	45,000

LAST_NAME	SSN	SALARY
Jaja	111-22-1111	40,000
Keenee	111-24-1245	60,000
Dizoza	111-87-2749	50,000
Jaja	111-49-2849	45,000

Figure 2.1 . Data Masking Example

Five steps Process for Basic Data masking Technique:

a. Upload and extract Production data into a common format.
b. Analyse inventory and classify Data.
c. Define Test Data creation rules.
d. Extract masked data subset.
e. Load into test environment.

3. COMPARISON

Till today many researcher has already developed many models and algorithm to address trad off between privacy and data utility. Let's understand some techniques and their pros and cons.

A. *Encryption*

In encryption, suppose data value X is replaced by h(X), where h is hash function. And no such h is provably known. This approach converts plain text in to cipher text. AES, SHA, etc. algorithms are used in encryption. For, Example data value of balance" 2000" is encrypted to "A3tv5".

Advantages

Encryption technique is easy to implement and it is efficient. This approach is useful for simple setting.

Disadvantages

This technique provides privacy but it lost the richness of original data. So, data utility is very low. It is poor at discovery of Data Knowledge. Moreover, when domain is small attacker try to dictionary attack and try all values of X.

B. *Perturbation*

Another approach data Perturbation hides exact values of Data. For example adding noise to data and its numerous improvements. In this approach richness of original data is captured. Such technique is found very useful in Data mining. A simple perturbation technique can be include Gaussian noise as addition noise to data.

Suppose X is the input Data. So, result Y= X + e. Here e is Gaussian noise to the Data, which is taken from standard distribution of data as per shown in Figure 3.1.

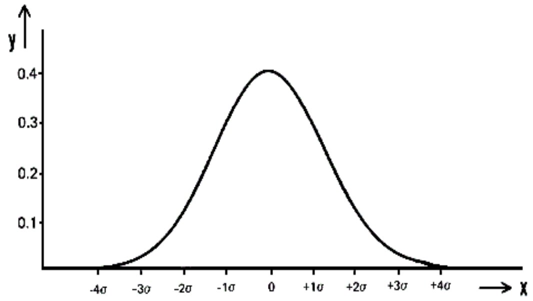


Figure.3.1 Gaussian Noise

Advantages

Perturbation approach is also very efficient and easy to implement. The analysis of this technique says that, this technique is capable of providing high Data utility and low disclosure risk.

Disadvantages

If anyone wants to draw interfaces with 100% confidence, this technique is not suitable [2].

C. *K-Anonymity*

K-Anonymity technique reduces the granularity of representation of these pseudo-identifiers with the use of techniques such as generalization and suppression [8]. In the method of generalization, the attribute values are generalized to a range in order to reduce granularity of representation. And in the method of suppression, the attribute values are removed completely. For example, the date of birth could be generalized to a range such as year of birth. So, in this technique we guaranteed 50% of privacy.

Advantages

This technique provides high Data utility. Actually the data which are sensitive are not change at all. This gives a solid mathematical bound on the identity disclosure risk.

Disadvantages

Disclosures of values are allowed. Since many of the attributes are unperturbed and so true values are exposed. It is very hard to do optimally and good approximation may not be possible.

A comparative overview of all above technique is below:

Table.1 Comparison

	Data Utility	**Identity Disclosure Risk**	**Value Disclosure Risk**
Encryption	Low	Medium	Medium
Perturbation	Medium	Low	Low
k-Anonymity	High	Low	High

As per this Comparison table none of these techniques are very satisfactory. So there is a need to achieve both Data Utility and High privacy with a single Stroke. Data masking technique try to solve this problem.

4. EXISTING DATA MASKING TECHNIQUES

One of leading research group TRDDC of TCS comes with some data Masking Solutions. They use these masking techniques to protect privacy for real time Data. Some existing data masking solutions of TRDDC group are below.

A. Masketeer

The objective of this tool I still de-identify data or exclude personal identity from Data. And still providing high Data utility and Privacy.

This TCS tool comes for standalone client application. And ensure to protect Production data. This tool is Platform and Database independent. MASKETEER^{TM} [9] has a whole bug of different techniques providing mathematical robustness guarantee for masking purposes. MASKETEER^{TM} is available on mainframe, Windows, Linux, MAC and UNIX. This TCS tool is very user friendly. MASKETEER^{TM} mainly introduce for static data and also known as offline masking solution.

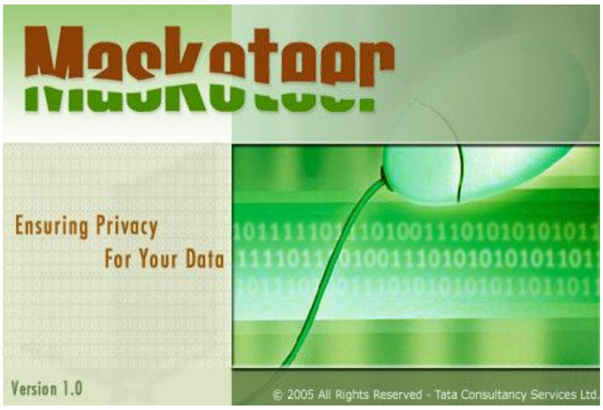


Figure.4.1 MASKETEER^{TM}

B. Safe Mask

TRDDC group comes with this Safe Mask tool for online masking of Data. This data masking tool allows people to see only *need-to-know* data online.

Safe mask architecture contains two main components: (1) Privacy Aid Station, and (2) Privacy Capsule.

Privacy Aid Station is for creating and maintaining privacy rules. And it acts as central control and provides support for the role based privacy.

Privacy Capsule has no rule inside. It fetches rules from Privacy Aid Station and as per rules it prescribed application URLs.

Architecture of safe mask as per TRDDC group is shown in Figure 4.2.

Here Privacy capsule has main role as it masks the Data. Privacy Capsule follows the rules of Privacy Aid Station. Rules can be modified later as per requirement.

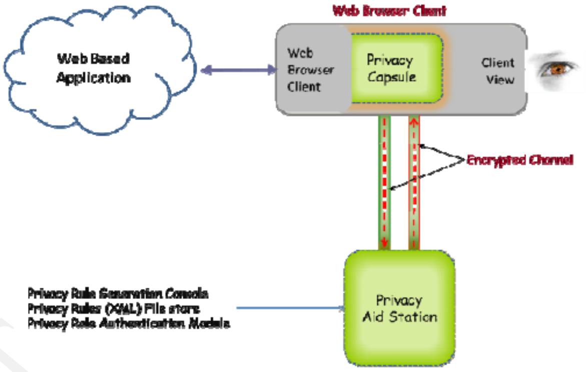


Figure.4.2 Safe Mask Architecture

5. CONCLUSION

I can summarize my final conclusion from my analysis as Data Masking, new immerging technique for data privacy preservation is capable of fulfill requirements, which other techniques are not able to provide. Data masking technique is easy to implement. It provides high Data Utility and High Privacy Preservation as compare to other techniques. Groups like TRDDC of TCS are currently working to improve Data masking technique.

6. REFERENCES

[1] Wherever Gartner (1999) Inc. Server Storage and RAID Worldwide, Technical report, Gartner group. Available: http://www.gartner.com

[2] Sachin Lodha and Sharda Sundaram, "Data Privacy," *TACTiCS – TCS Technical Architects' Conference'05,* 2005.

[3] Ravikumar G K, Manjunath T N, Ravindra S Hegadi and Umesh I M, "A Survey on Recent Trends, Process and Development in Data Masking for Testing," *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 2, pp. 535-544, March. 2011.

[4] Muralidhar, K. and R. Sarathy," A Theoretical Basis for Perturbation Methods," *Statistics and Computing,*Volume 13, Issue 4 , pp. 329-335, 2003.

[5] Sarathy, R., K. Muralidhar, and R. Parsa," erturbing Non-Normal Confidential Attributes: The Copula Approach," *Management Science*, 48(12), 1613-1627, 2002.

[6] Muralidhar, K. and R. Sarathy," A Rejoinder to the Comments by Polettini and Stander on 'A Theoretical Basis for Perturbation Methods'," *Statistics and Computing*, 13(4), 339-342, 2003.

[7] Mini Li, Zheli Liu, Chunfu Jia and Zongqing Dong, "Data masking generic Model," *IEEE ,Fourth International Conference on Emerging Intelligent Data and Web Technologies,* 2013

[8] Samarati P, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001).

[9] Sachin Lodha, Nikhil Patwardhan, AshimRoy, Sharada Sundaram and Dilys Thomas, "Data Privacy Using MASKETEER^{TM}," *ICTAC 2012*, LNCS 7521, pp. 151–158, 2012.

[10] Sachin Lodha and Vijayanad Banahatti, "Safe Mask," *TACTiCS – TCS Technical Architects' Conference'09*, 2009.

[11] Ravikumar G K, Dr B Justus Rabi and Manjunath TN, "A Study on dynamic data Masking with its Trends and implementations," *International Journal of Computer Applications (0975 – 8887)* ,Volume 38– No.6, January 2012.

[12] S.Vijayarani and Dr.A.Tamilarasi, "An Efficient Masking Technique for Sensitive Data Protection," *IEEE-International Conference on Recent Trends in Information Technology, ICRTIT* ,2011